

Long-term reliability and observer comparisons in the radiographic diagnosis of periapical disease

O. Molven¹, A. Halse², I. Fristad¹

¹Department of Odontology, Endodontics, and ²Radiology, School of Dentistry, University of Bergen, Norway

Abstract

Molven O, Halse A, Fristad I. Long-term reliability and observer comparisons in the radiographic diagnosis of periapical disease. *International Endodontic Journal*, **35**, 142–147, 2002.

Aim The aim of this study was to evaluate and compare the long-term diagnostic consistency of two examiners, an endodontist and a radiologist, and to make comparisons with findings recorded by an observer with more recent scientific and clinical experience in endodontics.

Methodology Three groups, each consisting of 20 full mouth series of intraoral radiographs, with 79, 93 and 85 endodontically-treated roots, respectively, were successively evaluated for periapical disease. Evaluations were at first performed separately by the three observers. Disagreement and difficult, borderline cases were subjected to joint evaluation. Intra- and interexaminer comparisons were made. For two of the observers the observations were compared with findings recorded

several years before for the same cases in the same radiographs.

Results The intra- and interobserver long-term reliability of the two original examiners resulted in 83% overall agreement, the kappa values were 0.54, 0.57 and 0.53. Comparisons between all three observers disclosed 82%, 85% and 86% agreement and kappa values 0.55, 0.58 and 0.60. The joint evaluations and decisions did not indicate a dominating influence from any of the observers.

Conclusions The long-term reliability of the two original observers was judged as being satisfactory. All three observers judged the overall disease status of the material in the same way. The joint discussions of selected cases might reduce observer variation to an acceptable level, avoid a number of false recordings and increase the reliability and validity of the findings.

Keywords: observers, periapical disease, radiographic diagnosis.

Received 11 June 2001; accepted 26 October 2001

Introduction

A strategy for the radiographic diagnosis of periapical pathosis was presented by Halse & Molven (1986) and used in follow-up studies, and they were later adopted by others (Halse & Molven 1987, Molven & Halse 1988, Sjögren *et al.* 1990, Saunders *et al.* 2000, Tronstad *et al.* 2000). This strategy involved two experienced observers, an endodontist and a radiologist. Cases were grouped either with no periapical pathological finding, with increased width of the periodontal ligament space, or with pathological finding. Agreement was studied on three levels: percentage agreement between scores,

agreement by calculation of Cohen's kappa, and discussed agreement, that is agreement after joint evaluation of disagreement and difficult, borderline cases. The use of this strategy indicated that: (a) the variation between the observers was reduced to an acceptable level; (b) obvious false recordings were few; and (c) diagnoses could be made which were directly related to the choice of treatment (Halse & Molven 1986).

The strategy has been reapplied by the same observers (OM and AH) in successive studies of treatment results now for the same root fillings 20–27 years postoperatively (Molven *et al.* 2002, unpublished observations). Another more recently qualified endodontist (IF) was introduced to the method, and it was decided to compare his observations with those made by the endodontist (OM) and the radiologist (AH). This was done in the present

Correspondence: Inge Fristad, Department of Odontology, Endodontics, School of Dentistry, University of Bergen, Årstadveien 17, N-5009 Bergen, Norway (fax: +47 55 58 66 30; e-mail: inge.fristad@odont.uib.no)

methodological study which primarily aimed at analyses of the long-term reliability of the original observers.

The purposes of this paper therefore are: (1) to present findings related to the long-term stability of two experienced observers and (2) to compare their observations with evaluations made by an observer with recent scientific and clinical training in endodontics.

Materials and methods

The material consisted of 60 full-mouth series of intraoral radiographs. The series, containing 257 endodontically-treated roots, had been taken at follow-up examinations 10–17 years after completion of the endodontic treatment in a teaching clinic, and had formerly been evaluated by two of the observers (OM, A and AH, B). The material was divided into three groups, each consisting

of 20 full-mouth series of radiographs, with 79, 93 and 85 endodontically filled roots, respectively.

The radiographic techniques and diagnostic procedures have been presented previously (Halse & Molven 1986). Three standard groups of findings were used (Figs 1–3). The evaluations were made by the two original observers and the new endodontist (IF, C) on three separate occasions. Each observer first evaluated one group containing 20 series of radiographs. Thereafter, a session of calibration and joint evaluation and decision (see later) followed before another group of radiographs was evaluated. A joint session also followed after evaluation of the third group of radiographs.

Calibration and decision procedure

Two observers' agreement was recorded as the radiographic result.



Figure 1 Normal periapical findings after endodontic treatment, schematically illustrated (left) and as observed in different regions of the jaws.

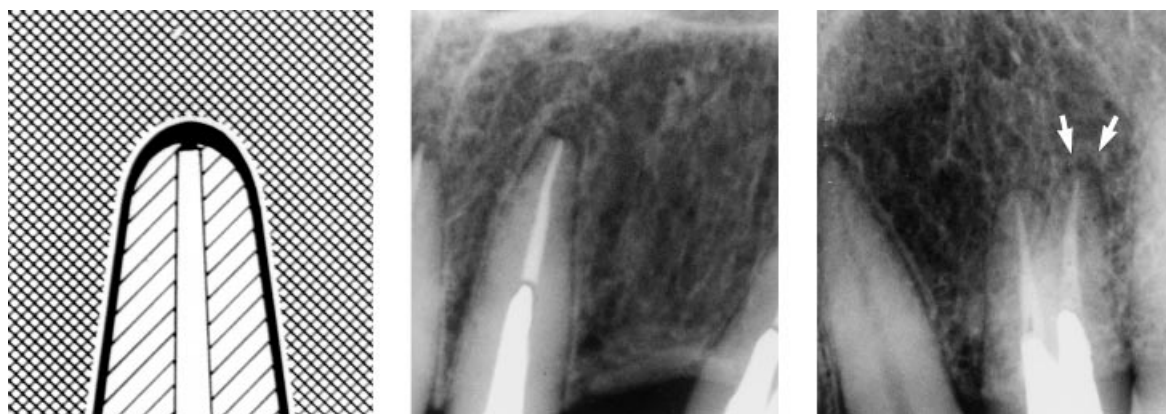


Figure 2 Widened periodontal spaces illustrated schematically (left) and as observed in different regions of the jaws. Note: The structure of the bone around the apex in the left radiograph was judged to be part of the normal trabecular system.

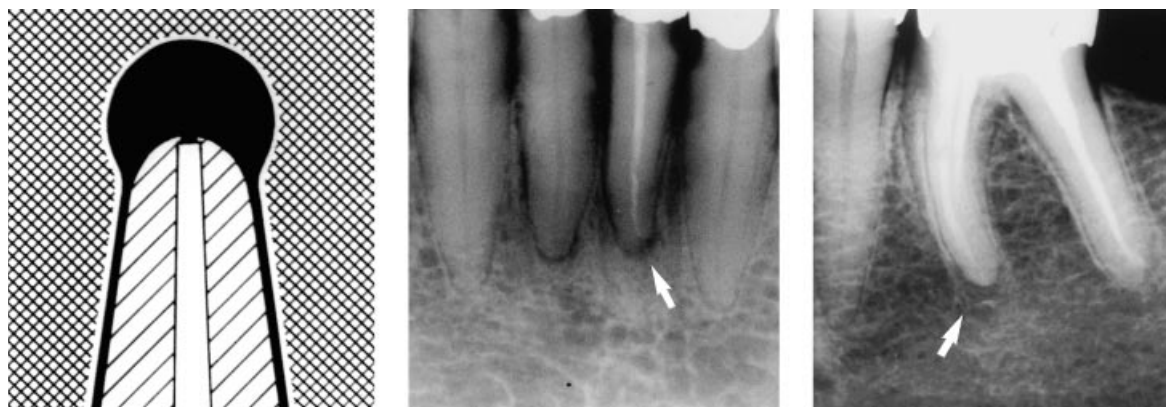


Figure 3 Pathological findings (periapical radiolucency) illustrated schematically (left) and as observed in different regions of the jaws.

Cases evaluated differently by the three observers were scheduled for joint discussion with an aim of consensus or majority decision. In addition to the calibration as a function of the joint evaluation of disagreement cases, some cases, suited for discussion, were selected by one of the endodontists (OM). They were also discussed and re-interpreted jointly immediately after each evaluating occasion at a meeting between the observers. Selectional guidelines were:

- i) each observer should be represented with deviations from the two others;
- ii) each classification group should be represented as a deviating diagnosis;
- iii) special attention should be given to difficulties encountered with the diagnosis of apical periodontitis; and
- iv) different tooth groups and both jaws should be included.

Rejection of radiographs

Radiographs rejected by the radiologist and one of the endodontists were omitted from the study. Radiographs rejected by the two endodontists were reevaluated by the radiologist to make a final decision about rejection.

Radiographs rejected only by the radiologist were subjected to joint evaluation.

Results

Long-term reliability

Two observers (A and B) had evaluated the same material 15 years earlier. Comparisons between earlier and present findings revealed 83% intraobserver agreement

Table 1 Periapical findings by three observers separately evaluating 257 endodontically treated roots, compared with the results after joint evaluation of disagreement cases and selected difficult, borderline cases. Results presented as percentages

Periapical findings	Observer			Joint evaluation
	A	B	C	
Normal width of the periodontal space	75.5	71.6	76.6	76.3
Increased width of the periodontal space	13.6	14.8	16.3	13.6
Pathological finding	6.6	10.1	6.6	7.8
Radiographs rejected	4.3	3.5	0.5	2.3

for both of them, with kappa values 0.54 and 0.57.

The corresponding interobserver figures were 83% and 0.53.

Observer comparisons

The observers' findings are presented in Table 1 together with the results after the joint evaluation of disagreement and difficult, borderline cases. Details regarding the latter cases are presented below.

Agreement between all observers was found for 73% of the roots.

The two original observers now had an interobserver agreement of 86%, kappa 0.61. The new endodontist's evaluation was close to those of the two original examiners. The agreement of A vs. C was 85%, kappa 0.58, and the agreement for B vs. C was 82%, kappa 0.55.

Disagreement and difficult borderline cases

A total of 32 cases (12%) were subjected to joint discussion. Three cases (1%) had been given different diagnoses

by the three observers. Eight rejections, either by the two endodontists (A and C) or by the radiologist (B) alone, were reevaluated. Twenty-one cases were selected as being suitable for discussion amongst the cases with initial agreement between the two observers.

Final agreement about the diagnoses was obtained for all cases except seven rejections that were maintained. The diagnoses for six of the 21 selected cases, with initial agreement between the two observers, were changed after discussion between all three observers. Analyses of the data did not indicate that any of the observers had a special influence on the joint decisions.

Discussion

The group of patients used in this study had been studied previously to determine changes in their periapical health status (Halse & Molven 1987, Molven & Halse 1988). In clinical situations, such observations form a basis for diagnostic conclusions regarding both overall and individual treatment results and therapeutic decisions. However, these data and conclusions are influenced by observer variations (Markén 1962, WHO 1997). The value of the findings therefore depends on a satisfactory observer performance and correspondence between the observers' judgement and what may be regarded as correct diagnoses (Koran 1976, WHO 1997, Wulff & Gøtzsche 2000).

In the present study each examiner, both the two original investigators and the one with a recent scientific and clinical training in endodontics, disclosed normal periapical conditions in approximately three out of four root-filled roots, periapical disease in 7–10% of the cases, and an increased width of the apical periodontal ligament space in the remaining cases (approx. 15%). Thus, they all judged the sample of endodontically treated roots to be characterized by a few teeth with pathosis and a high number of periapically healthy roots, a characteristic also maintained after the joint evaluation of disagreement and borderline difficult cases. Similar disease status has been reported in other follow-up samples of patients who have had root canal treatment in dental schools (Friedman 1998).

The observers' assessments of the overall disease status indicate a common opinion amongst two endodontists and the radiologist about the general disease status of the sample. The validity of this finding, however, has to be evaluated to judge its importance, and also because clinicians quite often overestimate their diagnostic competence and ability (Wulff & Gøtzsche 2000). Simultaneously,

information about the consistency of each of the three examiners and the variation between them is necessary for two reasons: (1) for revealing the long-term reliability of the original observers, and (2) for comparing their observations with evaluations made by the investigator more recently introduced to the diagnostic strategy.

Long-term reliability of original observers

The intraexaminer reproducibility, or each observer's long-term reliability calculated by comparing earlier and present observations, disclosed 83% agreement for both observers tested. Furthermore, when interobserver comparisons were made, the original investigation also revealed 83% agreement between the two examiners, whilst the present agreement was 86%. These findings indicate good intra- and interobserver agreement rates on both occasions. From a methodological point of view, they satisfy a general requirement that the percentage of agreement between scores should be in the range 85–98% (WHO 1997). Such levels of observer agreement are regarded almost as normal for the interpretation of radiographic images (Brorsson & Wall 1985) and this should be expected in samples with few periapical pathoses, probably reflecting the observers training and experience and the quality of the images. When the prevalence of disease is low, the figures should be calculated to show levels of reproducibility above those expected to occur by chance (Koran 1976, Bulman & Osborn 1989, Wulff & Gøtzsche 2000). The kappa statistic gives such figures and is a more valid assessment of intra- and interobserver agreement compared to the percentage of agreement between scores. The present kappa values, from 0.53 to 0.61, i.e. true agreement levels from 53% to 61%, are regarded as good ratings for evaluation of skeletal structures (Cockshott & Park 1983). Corresponding values have been disclosed in other endodontic investigations (Trope *et al.* 1999, Saunders *et al.* 2000), and higher values, indicating 80% corrected agreement or more have also been presented (Sjögren *et al.* 1990, Weiger *et al.* 1997, Kirkevang *et al.* 2000). Differences pertaining to the number of diagnostic groups and the frequency of diagnoses, may explain the latter values if compared with the present ones. Therefore, it is reasonable and relevant to conclude that the long-term reliability of the two original observers was good with a moderate to substantial agreement between the present observations and findings made several years earlier, for the same cases viewed on the same series of radiographs.

Original observers vs. new examiner

Long-term follow-up studies often imply that observers are brought in for practical, methodological and also educational purposes. These examiners must be tested against standard requirements of observer judgements, and compared with the performance of so-called experts or more experienced observers. The interpretation, understanding and application of codes and criteria should be uniform (Koran 1976, WHO 1997, Wulff & Göttsche 2000). Each observer should examine consistently, and original observers and others more recently introduced to the method should be closely correlated in their judgements.

The present findings indicate that these requirements were fulfilled. The interobserver agreement was above 80%, and the kappa values 0.55, 0.58 and 0.61 revealed good reproducibility. Thus, judgements made by the observer with a more recent scientific and clinical training in endodontics corresponded to those made by the two original observers. The three observers therefore appeared to interpret radiographs in the same way, indicating that they were calibrated against a standard resulting in observations with no marked influence from bias and systematic error (Halse & Molven 1986).

Joint agreement

Observer error and bias is part of clinical research, and can never be eliminated (Koran 1976, WHO 1997, Wulff & Göttsche 2000). Measures must, however, be taken to minimize their effect. Therefore, in studies of treatment results after conventional root canal filling and after endodontic surgery, the importance of joint evaluations as part of the diagnostic strategies has been emphasized (Halse & Molven 1986, Molven *et al.* 1987). Thorough discussions before deciding about cases recorded as being difficult (that is borderline and deviating cases identified during the investigation) would be expected to increase the chances of obtaining reliable and valid radiographic data. Joint discussions during the study should also ensure that the classification system is continuously repeated and discussed in relation to diagnostic problems, and a calibration effect is likely to be expected. By these measures the risk of serious observer deviations and obvious wrong recordings should be reduced to an acceptable minimum.

In the present study we included three occasions for discussed agreement, one after each separate evaluation of one third of the material. Altogether 12% of the material was subjected to joint discussions and a decision was

obtained for all the reevaluated cases including seven rejections. In an earlier investigation by just two of the same observers, about 18% of the material was scheduled for joint evaluation (Molven & Halse 1988). This suggests that several difficult cases can be observed even in samples with a presumably great number of easily detectable normal findings. Comparable figures are not readily found in the literature and should be given to illustrate diagnostic difficulties in studies otherwise satisfying general methodological requirements regarding observer reproducibility.

The diagnostic conclusions in the difficult cases, the disease or no disease decisions, are important for the estimation of the overall success percentages. And, as also discussed by Kvist (2001), they are crucial as a basis for therapeutic decisions in individual cases.

Conclusions

The long-term stability of the two original observers was satisfactory, with a similar level of intraexaminer agreement as the original figures.

The observer more recently introduced to the method for the diagnosis of periapical pathosis made similar judgements of the overall disease status of the material.

The joint evaluation of disagreement and difficult borderline cases is expected to reduce observer variation to an acceptable level and to avoid a number of false recordings.

References

- Brorsson B, Wall S (1985) Validity and deduction. Assessment of Medical Technology – Problems and Methods. Swedish Medical Research Council, Stockholm, 34.
- Bulman JS, Osborn JF (1989) Measuring diagnostic consistency. *British Dental Journal* **166**, 377–81.
- Cockshott WP, Park WM (1983) Observer variation in skeletal radiology. *Skeletal Radiology* **10**, 86–90.
- Friedman S (1998) Treatment outcome and prognosis of endodontic therapy. In: Ørstavik, D, Pitt Ford, TR, eds. *Essential Endodontology Prevention and Treatment of Apical Periodontitis*. London, UK: Blackwell Science, 368–9.
- Halse A, Molven O (1986) A strategy for the diagnosis of periapical pathosis. *Journal of Endodontics* **12**, 534–8.
- Halse A, Molven O (1987) Overextended gutta-percha and Kloroperka. N-Ö root canal fillings. Radiographic findings after 10–17 years. *Acta Odontologica Scandinavica* **45**, 171–7.
- Kirkevang L-L, Ørstavik D, Hörsted-Bindslev P, Wenzel A (2000) Periapical status and quality of root fillings and coronal restorations in a Danish population. *International Endodontic Journal* **33**, 509–15.

- Koran LM (1976) Increasing the reliability of clinical data and judgments. *Annals of Clinical Research* **8**, 69–73.
- Kvist T (2001) Endodontic Retreatment. Aspects of decision making and clinical outcome. Thesis, Göteborg, Sweden. *Swedish Dental Journal* (Suppl. 144).
- Markén K-E (1962) Studies of deviations between observers in clinico-odontological recording. Thesis. Uppsala, Sweden: Almqvist & Wiksells Boktryckeri AB.
- Molven O, Halse A (1988) Success rates for gutta-percha and Kloroperka N-Ö root canal fillings made by undergraduate students: radiographic findings after 10–17 years. *International Endodontic Journal* **21**, 243–50.
- Molven O, Halse A, Grung B (1987) Observer strategy and the radiographic classification of healing after endodontic surgery. *International Journal of Oral Maxillofacial Surgery* **16**, 432–9.
- Saunders MB, Gulabivala R, Holt R, Kahan RS (2000) Reliability of radiographic observations recorded on a proforma measured using inter- and intra-observer variation: a preliminary study. *International Endodontic Journal* **33**, 272–8.
- Sjögren U, Hägglund B, Sundqvist G, Wing K (1990) Factors affecting the long-term results of endodontic treatment. *Journal of Endodontics* **16**, 498–504.
- Tronstad L, Asbjørnsen K, Døving L, Pedersen I, Eriksen HM (2000) Influence of coronal restorations on the periapical health of endodontically treated teeth. *Endodontics and Dental Traumatology* **16**, 218–21.
- Trope M, Olutayo Delano E, Ørstavik D (1999) Endodontic treatment of teeth with apical periodontitis: Single vs multivisit treatment. *Journal of Endodontics* **25**, 345–50.
- Weiger R, Hitzler S, Hermle G, Löst C (1997) Periapical status, quality of root canal fillings and estimated endodontic treatment needs in an urban German population. *Endodontics and Dental Traumatology* **13**, 69–74.
- World Health Organization (1997) *Oral Health Surveys, Basic Methods*, 4th edn. WHO, Geneva, Switzerland: 13–5, 62–3.
- Wulff HR, Göttsche PC (2000) *Rational Diagnosis and Treatment. Evidence-Based Clinical Decision-Making*. London, UK: Blackwell Science, 29.